Advances in Social Science, Education and Humanities Research, volume 233
3rd International Conference on Contemporary Education, Social Sciences and Humanities (ICCESSH 2018)

Quantitative Analysis of Research Tendency on Hotspot of Library's Data Literacy Education in China

Yue Chen
Library
Jianghan University
Wuhan, China

***Abstract*—With the amount of papers from 2008-2017 year in China, which retrieved by terms of "data literacy education" and "library", this paper applies metabolic GM(1,1) forecasting model to explore variation trend of the hot-spot in the big data era. The result indicates that the proportion of paper's amount based on "data literacy education" to total refers to "library", keeps ever-increasing in the ration of 39.22% since 2015.**

Keywords—data literacy education; gray system theory; metabolic GM(1,1)

I. INTRODUCTION

The ability to construct knowledge from data is displayed more significantly in data literacy than that of traditional information literacy. In the ear of big data, the library service will be significant changes in functions, data literacy education is one of the main functions of library in big data era [1][2]. Under the environment of big data, the important degree of data for the society gradually warms up, promotes the people's cognition of the personal literacy, as one of the education institutions, university library also begins to turn the traditional information literacy education to data literacy education [3][4].

The broad conception of "information literacy" is well known in librarianship. Many universities offer instruction in information literacy, either embedded throughout the curriculum or as a workshop series through the library system. Much has been written about the importance of teaching students the ability to find and evaluate sources of information. When students are able to "determine their own information needs, to use some information retrieval tools efficiently, to evaluate the retrieved information, and to use that information to answer their needs" [5], students feel more confident in their abilities and less anxious, make better use of the library resources available to them, and learn transferable literacy skills.

As we move towards a world run by data it is essential that we increase data literacy to help people stay informed, and decisions to be made based on reasoned fact rather than somebody else' interpretation.

What has changed in the information landscape is the scope and depth of data available. Data literacy education is also a relatively new concept to librarianship.

This paper took paper amounts which research on "data literacy education" and "library" as analysis and modeling objects. Firstly, took Knowledge Tendency Analysis on the hot-spot; secondly, took full text literacy database as data source, acquired and analyzed the data since 2008 year by Excel; thirdly, applied gray system theory [6][7] to explore transformation tendency of the hot-spot in the big data era; and summarized the explored research tendency which on the relevant hot-spot.

II. HIT RATE ANALYSIS OF THE HOT-SPOT

Entered the interface of "Knowledge Trend Analysis" by logging in the Wanfang data platform, we took "data literacy education" as search term, so the hit term could be got from each year during 2014-2017. In this way, we could find out the transformation tendency of the relevant hot-spot with the hit rates, and pave the way for the quantitative analysis and trend mining. The transformation process was showed in "Fig. 1".

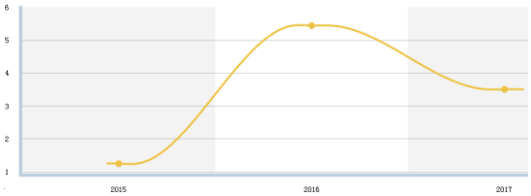

Fig. 1. Knowledge trend analysis curve of hotspot on "data literacy education" during 2014-2017.

The paper took CNKI as data source, the sample span was from 2008 to 2017 year. Curve of "Data Literacy Education" and "Library" relevant paper quantities during 2008-2017 was showed in figure 2. Taking statistics with these data by Excel, we analyzed the proportion of paper's amount based on "data literacy education" to total refers to "library", statistical results were showed in "Fig. 3".

With the results searched out, research on "data literacy education of library" during 2008-2010 is nearly none. So, in

order to explore the relevant research tendency, we select historical data of 2011-2017 years as modeling objects.

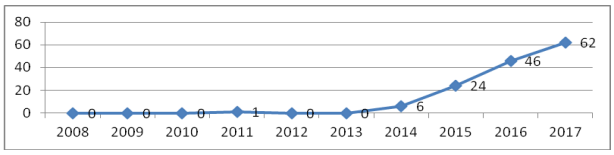

Fig. 2. Paper quantities of "data literacy education" and "library" during 2008-2017.

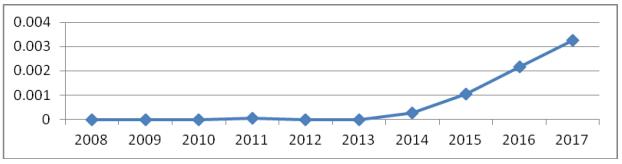

Fig. 3. The proportion of paper's amount based on "data literacy education" to total refers to "library" during 2008-2017.

III. MODELING WITH GRAY SYSTEM THEORY

In the natural and social systems, the problems of uncertainty was widespread. Systems with lots of samples can be analyzed by probability and statistics. In other cases, fuzzy mathematics can be used to solve the problem. However, due to the lack of samples, poor information and lack of experience, there are also problems with uncertainty, which can be described in gray model(GM), when the processes are complex and indescribable with precision and accuracy through the use of mathematical models. The gray system theory was originally proposed by Chinese researcher Deng (1982), and developed by Liu.

According to the central limit theorem, probability theory and mathematical statistics which with large sample data, the traditional statistical model needs to meet the requirements of large sample. However, due to accelerated development of the IT, academic research which on data literacy education of library was proposed on the last decade. In practical applications, the traditional mathematical statistical model has certain limitations to the amount of samples, but fortunately, gray theory model requires only a small number of samples, simple calculation, but have higher adaptability and is more reliable. In this way, the gray theory makes up for the problem of fewer samples in the modeling process under the conditions of "small sample and poor information". For this reason, the approach based on gray system theory is appropriate for forecasting the proportion of paper's amount based on "data literacy education of library" to total refers to "library".

The processes of proportion prediction with GM(1,1) model are presented as follows:

- Assume that $x^{(0)} = \{x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n)\}$, which collected from the historical proportion of the first n years, is a given sequence of raw data.
- Applying 1-AGO(Accumulated generating operation) on $x^{(0)}$, provided that $x^{(1)} = \{x^{(1)}(1), x^{(1)}(2), \cdots, x^{(1)}(n)\}$, where $x^{(1)}(i) = \sum_{k=1}^{i} x^{(0)}(k), i = 1, 2 \cdots, n$.
- Apply a consecutive neighbor mean generation to $x^{(1)}$. Let $z^{(1)}(k) = \frac{1}{2}x^{(1)}(k-1) + \frac{1}{2}x^{(1)}(k), k = 2, 3, \cdots, n$, then it follows that $z^{(1)} = \{z^{(1)}(2),\ z^{(1)}(3),\ \cdots,\ z^{(1)}(n)\}$.
- Construct a white differential equation based on gray system theory: $\frac{dx^{(1)}}{dt} + ax^{(1)} = u$, where a means the developing coefficient, u the gray input. Solving the equation, the model response can be given by

$$x^{(1)}(t) = (x^{(1)}(t_0) - \frac{u}{a})e^{-a(t-t_0)} + \frac{u}{a},$$

- where t_0 is the initial time moment. If we sample the equal-time-interval discretely, the time response obtained, which follows that

$$x^{(1)}(k'+1) = (x^{(1)}(1) - \frac{u}{a})e^{-ak'} + \frac{u}{a},$$

- where k' is the sampled time moment, and valued as positive integer from 1.
- Perform a least squares estimate for the parameters (a and u). The simulated parameters ($\hat{a}$、$\hat{u}$) can be obtained that

$$\begin{bmatrix} \hat{a} \\ \hat{u} \end{bmatrix} = \left(B^T B\right)^{-1} B^T Y_n,$$

where

$$B = \begin{pmatrix} -z^{(1)}(2), & 1 \\ -z^{(1)}(3), & 1 \\ \cdots & \cdots \\ -z^{(1)}(n), & 1 \end{pmatrix},$$

$$Y_n = (x^{(0)}(2),\ x^{(0)}(3),\ \ldots\ ,\ x^{(0)}(n))^T.$$

At this time,

  - If $|\hat{a}| \le 0.3$, the relative mid or long term forecasting can be conducted.
  - If $0.3 < |\hat{a}| \le 0.5$, the short term prediction is preferred.
  - If $0.5 < |\hat{a}| < 0.8$, put the $x^{(1)}$ value back to $x^{(0)}$ value, then back to the step 2).
  - If $|\hat{a}| \ge 0.8$, should update the raw data, the smoothing process can be used here for example, then back to the step 2).
- Put the simulated parameters ($\hat{a}$、$\hat{u}$) back to the differential equation in step 4), the response is

$\hat{x}^{(1)}(k'+1) = (x^{(1)}(1) - \frac{\hat{u}}{\hat{a}})e^{-\hat{a}k'} + \frac{\hat{u}}{\hat{a}}$, and $\hat{x}^{(1)}(1) = x^{(0)}(1)$.

When the time with $k' = 1,2,\text{L}\ ,n-1$, the estimated sequence is called a simulated value of gray model;

When the time with $k' \geq n$, the estimated sequence is called a forecasting value of gray model.

- Evaluate the errors and precision. Assume that $\varepsilon(k')$ is the relative error, $\overline{\varepsilon}$ is the average relative error, and the precision is τ:

$\varepsilon(k') = \frac{x^{(1)}(k') - \hat{x}^{(1)}(k')}{x^{(1)}(k')} \times 100\%$, $\overline{\varepsilon} = \frac{1}{n-1}\sum_{k'=2}^{n}|\varepsilon(k')|$, $\tau = 1 - \overline{\varepsilon}$.

- Examine the evaluated precision, if less than the pre-set threshold (in the other words, it's not meeting a predetermined requirement), it needs to reduce possible errors, which caused in reciprocating operations, by constructing the remnant GM(1,1) model. The detail process follows that establish a GM(1,1) model using the error sequence

$\varepsilon'^{(1)}(k') = x^{(1)}(k') - \hat{x}^{(1)}(k')$

- firstly, and then add the estimated sequence $\hat{\varepsilon}'^{(1)}$ of remnant GM(1,1) model to the $\hat{x}^{(1)}$, repeat the modification process, until the precision meet the requirement.

- Restore the final estimated sequence through inverse accumulating (IAGO), the number of IAGO times should be equal to the AGO times. Therefore,

$\hat{x}^{(0)}(k') = \hat{x}^{(1)}(k') - \hat{x}^{(1)}(k'-1)$

is the sequence of proportion predicted (Y_e).

- Construct a metabolic GM(1,1) model, which built on the following new sequence

$x^{(0)} = \{x^{(0)}(2),\text{L}\ ,x^{(0)}(n), x^{(0)}(n+1)\}$

obtained by inserting $x^{(0)}(n+1)$ and deleting $x^{(0)}(1)$. As a matter of fact, as time goes on, some stochastic interferences or driving forces are concerned with the development of gray system once a metabolic GM(1,1) model established, therefore, the high accuracy can be achieved.

In the primitive GM(1,1) model modeling, the past data from the real time $t = n$ is used. However, the development of any gray system, as time goes on, will continue to have some random disturbance factors into the system, the progression of the system impacted. Therefore, with the primitive GM(1,1) model, the higher accuracy is only with a few recent data, deviate from reality, poorer effect of the prediction.

In order to impair the disturbance of the future random disturbance on gray system and improve the forecasting accuracy, the GM(1,1) model is reformed.

In the original data $x^{(0)} = \{x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n)\}$, the latest information $x^{(0)}(n+1)$ is placed and the oldest data $x^{(0)}(1)$ and $x^{(0)} = \{x^{(0)}(2), x^{(0)}(3),..., x^{(0)}(n+1)\}$ is removed. Same with the above steps, the model gray metabolic GM (1,1) model is established, and a series of forecasting data is inferred out, at the same time, the precision of metabolic GM(1,1) is higher than that of primitive GM(1,1) model.

IV. CONCLUSION

Due to the late start of the academic research on library's data literacy education, it can be seen from figure 1 and figure 2 that the research on data literacy education of library has rapidly developed in the last five years, so we have few numbers of the related papers.

In this way, there's only gray model fits for the prediction which requires less sample, simple calculation, and makes up the problems in paper quantity modeling under "little sample and poor information" conditions.

Obtained by statistic, proportion of paper's amount based on "data literacy education" and "library" to total refers to "library" each year during 2008-2010 is zero. So, we take GM(1,1) proportion modeling with historical data of 2011-2017 years, the average simulation error up to 72%, obviously, it does not meet the conventional requirements of precision. Considering with theorem of "recent data is most useful" from gray system theory, we could establish a metabolic GM(1,1) model, metabolize the last data successively. In this way, the average simulation error is 50.21% when modeling with historical data of 2012-2017 years; the average simulation error is 21.5% when modeling with historical data of 2013-2017 years; the average simulation error is 8.5% when modeling with historical data of 2014-2017 years; the average simulation error is 1.85% when modeling with historical data of 2015-2017 years.

In order to improve accuracy, we finally selected historical data from 2015-2017 years, coefficient vector of differential equation was calculated out as

$\hat{a} = [a,u]^T = [-0.392197\ ,0.001344\ ]^T$, and time response function of differential equation is

$\hat{X}^{(0)}(k+1) = 0.004469\ \exp(0.392197\ *k) - 0.003427$.

There were 3 sampled data, so we took 2-step forecasting, the predicted results were 0.004702 and 0.006959. In other words, proportion of paper's amount based on "data literacy education" to total refers to "library" each year in 2018 and 2019 year would be 0.4702% and 0.6959%. According to the simulated and predicted results, we could find that proportion of "data literacy education of library" is taking a steady sustained growth, nearly at the rate of 39.22% since 2015 year.

In the rapid development of the big data era, it is necessary to improve the data literacy of the whole society. Improving contemporary college students' data literacy education is the era of big data requirements on universities, as significant teaching units academic libraries play a decisive role in this process. Data literacy education is conducive to the further development of college library's services in ability and contents.

ACKNOWLEDGEMENT

This research was sponsored by Wuhan Academy of Educational Science (the Important Project of "Thirteen five years" Municipal Education Science Programming NO. 2016A107).

REFERENCES

[1] Zhang Lulu, Zhang Qun, Kong Chengguo. A Literature Review of Researches on Data Literacy from the Perspective of Users, Document,Informaiton & Knowledge [J], 2018(1): 114-121

[2] Chen Zhang. The Library of the Era of Big Data and Data Literacy Education. Library and Information [J], 2014(4):117-119

[3] Zhang Wenliang, Liu Jingyi. Exploration of the Framework of Data Literacy Education System of Academic Library. Library Work in Colleges and Universities [J], 2017(4):80-84

[4] Deng Lijun, Yang Wenjian. Research on University Library Data Literacy Education under Big Data Environment. Library Development [J], 2016(1)

[5] Julien, H., & Boon, S. Assessing instructional outcomes in Canadian academic libraries. Library and Information Science Research [J], 2004, 26(2), 121

[6] Deng Julong. Grey Control System [M]. Wuhan:Huazhong Institute of Technology Press,1985

[7] Liu Sifeng,Xie Naiming. Gray System Theory and Its Applications [M].Beijing:Science Press, 2008(4).